RATES OF UNIFORM ALMOST-SURE CONVERGENCE FOR EMPIRICAL PROCESSES INDEXED BY UNBOUNDED CLASSES OF FUNCTIONS

by

David Pollard
Statistics Department
Yale University
August 1986

Abstract

Two uniform limit theorems are proved for empirical processes. The conditions of the theorems involve random covering numbers; for a Euclidean class of functions (a concept that generalizes the Vapnik-Cervonenkis property for classes of sets) these are readily checkable. The theorems are applied to problems in density estimation and non-parametric regression, leading to rates of uniform almost-sure convergence that are comparable to the best rates in the literature.

Research partially supported by NSF Grant No. DMS-8503347

Keywords and phrases: weighted empirical process; uniform limit theorem; random covering number; Vapnik-Cervonenkis class; Euclidean class; symmetrization; truncation; density estimation; non-parametric regression; cross-validation.

Running head: empirical processes

§1. Introduction

Many problems in non-parametric asymptotic theory can be reduced to applications of limit theorems for empirical processes indexed by classes of functions. Two such theorems are proved in this paper. They are applied (Section 4) to problems in density estimation and non-parametric regression, leading to sharp uniform almost-sure convergence results. The method of proof extends the well-known empirical process technique of symmetrization followed by a one-step approximation, which produces uniform exponential bounds involving random covering numbers. The novelty in the methods is the generalization of an idea introduced by Breiman et al. (1984, Chapter 12) to prove the empirical processes results needed for their tree-structured classification and regression procedures.

The theorems concern the behavior of the empirical measures $\{P_n\}$ generated by an independent sample $\xi_1,\xi_2,\ldots$ from a fixed probability measure P. For each P-integrable function f the difference $P_nf - Pf$ between the expected values with respect to P_n and P converges to zero almost surely. The theorems give weight functions, which depend on f, such that the weighted differences converge almost surely to zero, uniformly over classes $\mathcal{F}_n$. Informally, Theorem 2.1 gives conditions on $\mathcal{F}_n$ and a sequence of positive constants $\{\gamma_n\}$ that ensure $|P_nf - Pf|$ is eventually small compared with $P_n|f| + P|f| + \gamma_n$, for all f in $\mathcal{F}_n$. Theorem 2.4 replaces the $\mathcal{L}^1$ weight function by an analogous $\mathcal{L}^2$ weight function, $(P_nf)^{\frac{1}{2}} + (Pf^2)^{\frac{1}{2}} + \gamma_n$. In particular cases these strange weights simplify to produce uniform limit theorems for ratios of P_nf to Pf. For example, if all the functions are non-negative, and if either P_nf or Pf is bigger than γ_n for every f in $\mathcal{F}_n$, then Theorem 2.1 implies that $|P_nf - Pf|$ is uniformly small compared with $P_nf + Pf$, eventually. More formally, the theorem reduces to

(1.1) $$\sup_{\mathcal{F}_n} \left| \frac{P_n f}{Pf} - 1 \right| \to 0 \qquad \text{almost surely}$$

in this case. Precise statements of the theorems and a number of consequences of this type appear in Section 2.

The key concept underlying both theorems is that of a covering number. If $\mathcal{F}$ is a class of functions with envelope F (that is, $|f| \le F$ for all f in $\mathcal{F}$), and if Q is a measure on the space where the functions live, the $\mathcal{L}^1(Q)$ covering number $N_1(\varepsilon, Q, \mathcal{F})$ is defined as the smallest number of $\mathcal{L}^1(Q)$ balls of radius εQF, with centers in $\mathcal{F}$, needed to cover $\mathcal{F}$. (Note: This definition differs slightly from the definition in Section II.5 of Pollard (1984), but it agrees with the usage of Nolan & Pollard (1986).) That is, it is the smallest n for which there exists a collection $\mathcal{F}^*$ of n members of $\mathcal{F}$ such that: to each f in $\mathcal{F}$ there is an f^* in $\mathcal{F}^*$ with

$$Q|f - f^*| \le \varepsilon QF$$

Of course the definition has content only when $QF < \infty$. The $\mathcal{L}^2(Q)$ covering numbers $N_2(\varepsilon, Q, \mathcal{F})$ are defined by substitution of $\mathcal{L}^2(Q)$ norms for $\mathcal{L}^1(Q)$ norms; the corresponding inequalities for $\mathcal{F}^*$ are

$$Q|f - f^*|^2 \le \varepsilon^2 Q(F^2)$$

The notation does not record the dependence of both N_1 and N_2 on the envelope F. That disguises one subtlety in the role of F. It is not always desirable to choose F as $\sup_{\mathcal{F}}|f|$, the smallest envelope possible. Indeed one would usually want to make it as large as possible subject to some finiteness constraint on a moment $Pg(F)$. Theorem 2.1 shows why. Further discussion of the advantages of a large F appears in Section 3.

The theorems depend upon the existence of good bounds for the random covering numbers $N_1(\cdot, P_n, \mathcal{F}_n)$ and $N_2(\cdot, P_n, \mathcal{F}_n)$. Adequate bounds are most

readily available for those classes that Nolan and Pollard (1986) have dubbed Euclidean.

(1.2) Definition. A class $\mathcal{F}$ is said to be Euclidean(A,V) for the envelope F if for $0 < \varepsilon \leq 1$ there is a uniform bound on the covering numbers,

$$N_1(\varepsilon, Q, \mathcal{F}) \leq A\varepsilon^{-V} \qquad \square$$

If $\mathcal{F}$ is Euclidean there is a similar polynomial bound on its $\mathcal{L}^2$ covering numbers, because

$$N_2(\varepsilon, Q, \mathcal{F}) \leq N_1(\tfrac{1}{2}\varepsilon^2, Q_F, \mathcal{F}) ,$$

where Q_F denotes the measure with density F with respect to Q.

Nolan and Pollard (1986, Section 5) have listed several of the properties that make the concept tractable and useful. For example, if $K(\cdot)$ is a function of bounded variation on the real line then the class of all functions $f_{t,\sigma}(x) = K((t-x)/\sigma)$, with $t \in \mathbb{R}$ and $\sigma > 0$, is Euclidean for every envelope. This class and its higher dimensional analogues figure prominently in the applications to density estimation and non-parametric regression in Section 4. There is more about Euclidean classes in Section 3.

The rate of growth in the random covering numbers determines how fast the constants $\{\gamma_n\}$ can converge to zero. For example, if each $\mathcal{F}_n$ is Euclidean(A,V) and if $Pg(F) < \infty$ for a g satisfying mild conditions, then (1.1) will allow any $\{\gamma_n\}$ that decreases more slowly than $g^{-1}(n)n^{-1}\log n$. (Here $g^{-1}(n)$ refers to the inverse function, not the reciprocal of $g(n)$.) In the limiting case of a bounded F, where g could be made to increase arbitrarily rapidly, the $g^{-1}(n)$ factor drops away, leaving the familiar $n^{-1}\log n$ as the lower bound where uniform limit theorems begin to fail.

Questions of measurability are pushed into the background for most of the paper. A regularity condition known as permissibility takes care of the difficulties

that might arise from the manipulation of suprema over uncountable classes of functions. The words "permissible class" appear in the statement of results as a reminder of the need for some measurability condition, but precise definition of the concept and a discussion of why it is needed are postponed until Section 6.

§2. Statement of the results

The applications in Section 4 require the following theorems only for the special of Euclidean classes. Nevertheless the theorems are stated here in greater generality, in order to emphasize that the rate of convergence for $\mathcal{L}^1$ weightings is determined by $\mathcal{L}^1$ covering numbers, and for $\mathcal{L}^2$ weightings by $\mathcal{L}^2$ covering numbers. For easier comprehension of Theorem 2.1 the reader might restrict $\mathcal{F}_n$ to a single Euclidean class $\mathcal{F}$, and hold ε_n fixed. The series condition is then almost equivalent to: $n\gamma_n/g^{-1}(n)$ increases faster than $\log n$. For Theorem 2.4 think of ε_n^2 as a fixed large multiple of $n^{-1}\log n$ and γ_n decreasing like n^{-K} for some large constant K.

(2.1) Theorem

For each n let $\mathcal{F}_n$ be a permissible class with envelope F. Let $g(\cdot)$ be a non-negative increasing function for which

(i) $Pg(F) < \infty$

(ii) $g(x)/x$ is increasing for $x > 0$

If $\{\varepsilon_n\}$ and $\{\gamma_n\}$ are sequences of positive numbers for which

$$\sum_{n=1}^{\infty} \mathbb{P}[1 \wedge N_1(\varepsilon_n\gamma_n, P_n, \mathcal{F}_n)\exp(-n\varepsilon_n^2\gamma_n/g^{-1}(n))] < \infty$$

then

$$\mathbb{P}\{\sup_{\mathcal{F}_n} \frac{|P_nf - Pf|}{P_n|f| + P|f| + \gamma_n} > C\varepsilon_n \text{ infinitely often}\} = 0$$

for the constant $C = 8(1 + 3PF)$. □

(2.2) Corollary

For each n let $\mathcal{F}_n$ be a permissible class of non-negative functions that satisfies the conditions of Theorem 2.1. For C as above, if $2C\varepsilon_n \le \frac{1}{2}$ eventually then

$$\mathbb{P}\left\{\sup_{\mathcal{F}_n(\gamma_n)} \left|\frac{P_nf}{Pf} - 1\right| > 8C\varepsilon_n \text{ infinitely often}\right\} = 0$$

where $\mathcal{F}_n(\gamma_n) = \{f \in \mathcal{F}_n : P_nf + Pf \ge \gamma_n\}$. □

The first step in the proof of Theorem 2.1 will be a truncation of $f(\xi_i)$ to zero outside the set where $g(F(\xi_i)) \le i$. That will be where the factor $g^{-1}(n)$ comes from. If F is bounded the truncation is superfluous; the $g^{-1}(n)$ can be discarded. For ease of later application the next Corollary records this fact in a slightly modified form.

(2.3) Corollary

For each n let $\mathcal{F}_n$ be a permissible class with constant envelope α_n. If $\{\varepsilon_n\}$ and $\{\gamma_n\}$ are sequences of positive numbers for which

$$\sum_{n=1}^{\infty} \mathbb{P}[1 \wedge N_1(\varepsilon_n\gamma_n/\alpha_n, P_n, \mathcal{F}_n)\exp(-n\varepsilon_n^2\gamma_n/\alpha_n)] < \infty$$

then

$$\mathbb{P}\left\{\sup_{\mathcal{F}_n} \frac{|P_nf - Pf|}{P_n|f| + P|f| + \gamma_n} > 32\varepsilon_n \quad \text{infinitely often}\right\} = 0 \qquad \square$$

Theorem 2.1 and its corollaries work best for functions with $Pf \approx \gamma_n$. For larger values of Pf the natural L^2 weighting seems to give a better bound.

(2.4) Theorem

For each n let $\mathcal{F}_n$ be a permissible class with envelope F for which $PF^2 < \infty$. If $\{\varepsilon_n\}$ and $\{\gamma_n\}$ are sequences of positive numbers such that

$$\sum_{n=1}^{\infty} \mathbb{P}[1 \wedge N_2(\varepsilon_n\gamma_n, P_n, \mathcal{F}_n)\exp(-n\varepsilon_n^2)] < \infty$$

then

$$\mathbb{P}\left\{\sup_{\mathcal{F}_n} \frac{|P_nf - Pf|}{(P_nf^2)^{1/2} + (Pf^2)^{1/2} + \gamma_n} > C\varepsilon_n \quad \text{infinitely often}\right\} = 0$$

where $C = 1 + 19(PF^2)^{1/2}$. $\square$

If $0 \le f \le 1$ for all f the assertion of the theorem is close to the convergence results obtainable by means of Le Cam's (1983) square-root technique. Indeed, that technique works because it manages to conjure up an extra factor like $(P_nf^2)^{1/2}$ (see the argument leading up to line 3 on page 33 of Pollard (1984)).

When specialized to uniformly bounded Euclidean classes the theorem does not quite include all the results of Alexander (1985). His Corollaries 1.6 and 1.7 work for γ_n down near n^{-1}, whereas my results need it to decrease no faster than $n^{-1}\log n$. This reflects the crudeness of the one-step approximation that will be used in my proofs to get the maximal inequality. It seems that improvement could come only by means of a chaining argument, which is the technique that produced Alexander's inequalities.

The assertion of the Theorem 2.4 would be more satisfactory if the factor $(P_nf^2)^{\frac{1}{2}}$ were not present in the denominator. Intuitively, it should be possible to absorb it into the other terms in the denominator because one would expect $P_nf^2 \approx Pf^2$ if Pf^2 were not too small. A rigorous argument to this effect may be based on Theorem 2.1 applied to the classes

$$\mathcal{F}_n^2 = \{f^2 : f \in \mathcal{F}_n\}$$

with envelope F^2. The $\mathcal{L}^1(Q)$ covering numbers for $\mathcal{F}_n^2$ are related to the $\mathcal{L}^2(Q)$ covering numbers for $\mathcal{F}_n$ by the inequality

$$N_1(2\varepsilon, Q, \mathcal{F}_n^2) \le N_2(\varepsilon, Q, \mathcal{F}_n)$$

because

$$Q|f_1^2 - f_2^2| \le Q|f_1 - f_2|(|f_1| + |f_2|) \le (Q(f_1 - f_2)^2\, Q(4F^2))^{\frac{1}{2}}$$

To illustrate how Theorems 2.1 and 2.4 can be combined, suppose each $\mathcal{F}_n$ is Euclidean(A,V) for a constant envelope 1. Invoke Corollary 2.3 with $\varepsilon_n = 1/64$ and γ_n^2 equal to a suitably large multiple of $\lambda_n = n^{-1}\log n$. With probability one it is eventually true that

$$|P_nf^2 - Pf^2| \le \tfrac{1}{2}(P_nf^2 + Pf^2 + O(\lambda_n^{\frac{1}{2}})) \qquad \text{for all } f \text{ in } \mathcal{F}_n$$

which implies

$$P_nf^2 \le 3Pf^2 + O(\lambda_n^{\frac{1}{2}}) \qquad \text{for all } f \text{ in } \mathcal{F}_n$$

Combine this bound with the assertion of Theorem 2.4, for ε_n^2 and γ_n^2 equal to suitably large multiples of λ_n, to get

$$\sup_{\mathcal{F}_n} \frac{|P_n f - Pf|}{(Pf^2)^{1/2} + O(\lambda_n^{1/2})} = O(\lambda_n^{1/2}) \qquad \text{almost surely} \tag{2.5}$$

A special case shows how tight the bound is. Take P as the uniform distribution on $(0,1)$ and let each $\mathcal{F}_n$ consist of all indicator functions of intervals. With probability one, the longest interval containing no sample points has length of order λ_n. Take f equal to the indicator function of this interval to see that the supremum corresponding to the lefthand side of (2.5) is at least of order $\lambda_n^{1/2}$.

§3. Euclidean Classes

There are two properties that make the Euclidean concept worthwhile: they are stable under several algebraic and boolean operations; and there are simple criteria for identifying non-trivial Euclidean classes.

It is easy to show that the Euclidean property is preserved if the envelope is increased, and that it is inherited by subclasses. The property is also stable under the formation of pointwise sums, pointwise products, pointwise maxima, and pointwise minima. That is, if $\mathcal{F}$ is Euclidean for the envelope F and $\mathcal{G}$ is Euclidean for the envelope G then:

(i) $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is Euclidean for the envelope $F + G$;

(ii) $\{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is Euclidean for the envelope FG ;

(iii) $\{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$ and $\{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ are both Euclidean for the envelope $F \vee G$.

These are all straightforward consequences of inequalities for integrals. For example, to establish (ii) fix an $\varepsilon > 0$ and a measure Q . Define Q_F as the measure with density F with respect to Q . Define Q_G analogously. Let $\mathcal{F}^*$ and $\mathcal{G}^*$ be the finite subclasses for which

$$\min_{\mathcal{F}^*} Q_G|f - f^*| \le \varepsilon Q_G(F) \quad \text{for all } f \text{ in } \mathcal{F}$$

$$\min_{\mathcal{G}^*} Q_F|g - g^*| \le \varepsilon Q_F(G) \quad \text{for all } g \text{ in } \mathcal{G}$$

Then, for each f in $\mathcal{F}$ and g in $\mathcal{G}$ and their corresponding f^* and g^* ,

$$\begin{aligned} Q|fg - f^*g^*| &\le Q|f - f^*||g| + Q|g-g^*||f^*| \\ &\le Q_G|f-f^*| + Q_F|g-g^*| \\ &\le 2\varepsilon\, Q(FG) \end{aligned}$$

Thus

$$N_1(2\varepsilon,Q,\mathcal{F}\cdot\mathcal{G}) \le N_1(\varepsilon,Q_G,\mathcal{F})\; N_1(\varepsilon,Q_F,\mathcal{G})$$

where $\mathcal{F}\cdot\mathcal{F}$ is a temporary notation for the set of all products in (ii). The Euclidean property follows immediately. The arguments for (i) and (iii) are similar.

Nolan and Pollard (1986) have collected together some facts about covering numbers that imply other stability properties for Euclidean classes.

The main point of departure for the construction of Euclidean classes is a combinatorial property named after Vapnik and Cervonenkis. Recall that a class of subsets $\mathcal{C}$ of a set S is said to be a VC class (or a polynomial class in the terminology of Section II.4 of Pollard (1984)) if there exist constants A and V for which:

$$\text{card}\{C \cap E : C \in \mathcal{C}\} \le A[\text{card}(E)]^V$$

for every finite subset E of S. (Here card denotes cardinality.)

If f is a real-valued function on a set $\mathcal{X}$, define its graph as

$$\text{graph}(f) = \{(x,t) \in \mathcal{X} \otimes \mathbb{R} : 0 < t < f(x) \text{ or } 0 > t > f(x)\}$$

Trivial modifications of Lemma II.25 of Pollard (1984) establish the result: if $\{\text{graph}(f) : f \in \mathcal{F}\}$ is a VC class of sets, then $\mathcal{F}$ is Euclidean for every envelope. When specialized to the indicator functions of sets in a class $\mathcal{C}$, this result reduces to Lemma 7.13 of Dudley (1978), which may be reinterpreted as: the indicator functions of the sets in a VC class form a Euclidean class for the envelope 1. This property characterizes VC classes. [If $\mathcal{C}$ is not a VC class there exists for each n a set of n points, E_n, that is shattered by $\mathcal{C}$. For the probability measure Q_n that puts mass n^{-1} at each point of E_n, the indicators of all 2^n distinct sets of the form $C \cap E_n$ are at least n^{-1} apart in $\mathcal{L}^1(Q_n)$ distance; for a Euclidean class the number of sets this far apart should increase at a polynomial rate.] The equivalence identifies the Euclidean property as one natural generalization of the VC property from sets to functions. Dudley (1986) has investigated several other plausible generalizations.

If a non-constant envelope F is chosen for the class of indicator functions of sets in $\mathcal{C}$, the Euclidean property need no longer imply the VC property.

(3.1) Example

Let $\mathcal{X}$ be the set of non-negative integers and $\mathcal{C}$ be the class of all subsets of $\mathcal{X}$. Certainly $\mathcal{C}$ is not a VC class. But it is Euclidean for the envelope F define by $F(n) = 2^n$.

Given ε with $0 < \varepsilon \leq 1$ find the integer k for which $2^k \geq \varepsilon^{-1} > 2^{k-1}$. Let $\mathcal{C}^*$ consist of all 2^{k+1} subsets of $\{0,1,\ldots,k\}$. For each C in $\mathcal{C}$ define $C^* = C \cap \{0,1,\ldots,k\}$. Then for each measure Q on $\mathcal{X}$,

$$\begin{aligned} Q|C - C^*| &\leq Q[k+1,\infty) \\ &\leq 2^k \varepsilon \, Q[k,\infty) \\ &\leq \varepsilon \, QF \end{aligned}$$

Thus $N_1(\varepsilon,Q,\mathcal{C}) \leq 2^{k+1} \leq 4\varepsilon^{-1}$. □

The example shows how Euclidean properties can be forced on a bad class by choosing the envelope to be very large in a region of the underlying space where the class is badly behaved. Of course this artifice will succeed only if the sampling distribution P does not put too much mass in the bad regions; otherwise the moment condition $Pg(F) < \infty$ of Theorem 2.1, or $PF^2 < \infty$ of Theorem 2.4, would be violated.

§4. Applications and Examples

The three examples in this section all involve smoothing by means of a bounded kernel K. In each case I will assume that the class of all functions of the form $K\left(\frac{\cdot - t}{\sigma}\right)$ is Euclidean for its natural constant envelope. Conditions to guarantee this may be found in Section 5 of Nolan and Pollard (1986). When K satisfies the assumption, let us say that K is a Euclidean kernel. In one dimension every kernel with bounded variation is Euclidean.

(4.1) Example

Let $\xi_1, \xi_2, \ldots$ be independent observations from a distribution P on $\mathbb{R}^d$. Suppose P has a bounded density $p(\cdot)$ with respect to Lebesgue measure. Let K be a Euclidean kernel that integrates to 1. Without loss of generality (the positive and negative parts could be treated separately) assume K is non-negative. Write $f_{t,\sigma}(\cdot)$ for $K\left(\frac{\cdot - t}{\sigma}\right)$. For each $\sigma > 0$ define a density estimate

$$\hat{p}_n(t,\sigma) = (n\sigma^d)^{-1} \sum_{i=1}^{n} K\left(\frac{\xi_i - t}{\sigma}\right) = \sigma^{-d} P_n f_{t,\sigma}$$

The usual method of analysis first compares $\hat{p}_n(t,\sigma)$ with the smoothed density

$$p(t,\sigma) = P\hat{p}_n(t,\sigma) = \sigma^{-d} P f_{t,\sigma} .$$

Then deterministic arguments based on assumed smoothness of $p(\cdot)$ are used to handle the bias term. For example, if $p(\cdot)$ is uniformly continuous, then

$$\sup_{\sigma \le \alpha_n} \sup_t |p(t,\sigma) - p(t)| \to 0$$

for every sequence $\{\alpha_n\}$ that converges to zero. The results from Section 2 can handle the random component $\hat{p}_n(\cdot,\sigma) - p(\cdot,\sigma)$.

Let $\{\beta_n\}$ be a sequence of constants for which $n\beta_n^d / \log n \to \infty$. Fix $\varepsilon > 0$. Then, from Theorem 2.1, with probability one it is eventually true that for all t and σ,

$$|P_n f_{t,\sigma} - P f_{t,\sigma}| < \varepsilon(P_n f_{t,\sigma} + P f_{t,\sigma} + \beta_n^d)$$

This implies, for all σ with $\sigma \geq \beta_n$, and all t,

$$|\hat{p}_n(t,\sigma) - p(t,\sigma)| < \varepsilon[\hat{p}_n(t,\sigma) + p(t,\sigma)] + \varepsilon$$

or

$$-\frac{\varepsilon}{1+\varepsilon}(1 + 2p(t,\sigma)) < \hat{p}_n(t,\sigma) - p(t,\sigma) < \frac{\varepsilon}{1-\varepsilon}(1 + 2p(t,\sigma))$$

Because $p(\cdot)$ is bounded, we deduce that

$$\sup_{\sigma \geq \beta_n} \sup_t |\hat{p}_n(t,\sigma) - p(t,\sigma)| \to 0 \qquad \text{almost surely}$$

For a uniformly continuous $p(\cdot)$ this means that

$$\sup_{\beta_n \leq \sigma \leq \alpha_n} \sup_t |\hat{p}_n(t,\sigma) - p(t)| \to 0 \qquad \text{almost surely}$$

for every choice of α_n and β_n with $\beta_n \leq \alpha_n \to 0$ and $n\beta_n^d/\log n \to \infty$. Any σ_n lying in the band $[\beta_n, \alpha_n]$ produces a uniformly consistent estimate of the underlying density. The smoothing parameters could even be random and depend on t in some complicated fashion that produced awkward dependencies between the values $\sigma_n(t)$ and $\sigma_n(t')$. Provided

$$\mathbb{P}\{\beta_n \leq \sigma_n(t) \leq \alpha_n \quad \text{for all } t, \text{ eventually}\} = 1,$$

the corresponding sequence of density estimates would be uniformly consistent.

For example, suppose $\sigma_n(t)$ is chosen so that the closed ball $B_n(t)$ with radius $\sigma_n(t)$ and center t contains exactly $k(n)$ sample points. That is, $P_n B_n(t) = k(n)/n$. Assume that $K(x) = 0$ for $|x| > 1$, for otherwise the density estimate might not be even pointwise consistent in regions of zero density. Then $\hat{p}_n(t,\sigma_n(t))$ is uniformly consistent for a bounded, uniformly continuous $p(\cdot)$ if $k(n) \to 0$ and $k(n)/\log n \to \infty$. Write γ_n for $k(n)/n$. From Corollary 2.2 applied to the VC class of all closed balls in $\mathbb{R}^d$,

$$\sup_t \left| \frac{P B_n(t)}{\gamma_n} - 1 \right| \to 0 \qquad \text{almost surely} \tag{4.2}$$

Because the density $p(\cdot)$ is bounded, this implies that there exists a constant c for which, with probability one,

$$\inf_t \sigma_n(t)^d \geq c\gamma_n \qquad \text{eventually.}$$

The sequence $\beta_n = (c\gamma_n)^{1/d}$ is an appropriate choice. As the upper bound α_n choose numbers for which $\alpha_n \to 0$ and $\gamma_n = o(\alpha_n{}^d)$. Then

$$\sup_{\sigma_n(t)\leq\alpha_n} |\hat{p}_n(t,\sigma_n(t)) - p(t)| \to 0 \qquad \text{almost surely}$$

In the low density region where $\sigma_n(t) > \alpha_n$ there is simple bound:

$$\hat{p}_n(t, \sigma_n(t)) = O(\gamma_n/\sigma_n(t)^d) = o(1)$$

These two results imply the desired uniform convergence. For if $\varepsilon > 0$ then (4.2) implies that

$$\sup_{p(t)\geq\varepsilon} \sigma_n(t)^d = O(\gamma_n) = o(\alpha_n) \qquad \text{almost surely;}$$

the convergence of $\hat{p}_n(t,\sigma_n(t))$ to $p(t)$ is certainly uniform over the set $\{p \geq \varepsilon\}$. On the set $\{p < \varepsilon\}$, either $\sigma_n(t) \leq \alpha_n$, in which case the first result applies, or $\sigma_n(t) > \alpha_n$, in which case $\hat{p}$ and p are close because both are small. □

(4.3) Example

Let $\xi_i = (x_i,y_i)$ for $i=1,2,\ldots$ be independent observations from a distribution P on $\mathbb{R}^d \otimes \mathbb{R}$. Let $m(\cdot)$ denote the regression function of y_i on the random vector x_i; it is defined on the subset of $\mathbb{R}^d$ that supports the distribution of the x_i. Let P_n denote the empirical measure for the $\{\xi_i\}$.

For a Euclidean kernel K define two families of functions on $\mathbb{R}^{d+1}$ indexed by $\mathbb{R}^d \otimes (0,\infty)$:

$$f_{t,\sigma}(x,y) = K\left(\frac{x-t}{\sigma}\right)$$

$$g_{t,\sigma}(x,y) = yK\left(\frac{x-t}{\sigma}\right)$$

For simplicity suppose $|K| \le 1$. By assumption the class $\mathcal{F}$ of all $f_{t,\sigma}$ is Euclidean for the envelope 1. The class $\mathcal{G}$ of all $g_{t,\sigma}$ is Euclidean for the envelope $F(x,y) = |y|$, because

$$N_1(\varepsilon,Q,\mathcal{G}) \le N_1(\varepsilon,\mu,\mathcal{F})$$

where μ is the measure having density $|y|$ with respect to Q. Define estimates of the regression function for each $\sigma > 0$,

$$\hat{m}_n(t,\sigma) = \frac{P_n g_{t,\sigma}}{P_n f_{t,\sigma}}$$

The first step in the analysis of $\hat{m}_n$ compares it with the smoothed regression function

$$m(t,\sigma) = \frac{P g_{t,\sigma}}{P f_{t,\sigma}} ;$$

Then the second step treats the bias term $m(t,\sigma) - m(t)$. In this example I will consider only the first step, leaving to the reader the formulation of the smoothness assumptions needed to bound the bias term.

To avoid delicate problems with division by small numbers, restrict attention to a subset J of $\mathbb{R}^d$ for which

$$\liminf_{\sigma\to 0} \inf_{t\in J} \sigma^{-d} P f_{t,\sigma} > 0$$

$$\sup_{t\in J} (P|g_{t,\sigma}| + P|f_{t,\sigma}|) = O(\sigma^d) \quad \text{as} \quad \sigma \to 0$$

Such inequalities would follow from the conditions imposed by Mack and Silverman (1982); they also make uniform over J the pointwise inequalities justified by Stute (1986).

Suppose $P|y|^s < \infty$ for some $s > 1$. What conditions on $\{\beta_n\}$ will ensure that

$$\sup_{\sigma\ge\beta_n} \sup_{t\in J} |\hat{m}_n(t,\sigma) - m(t,\sigma)| \to 0 \quad \text{almost surely?} \tag{4.4}$$

When coupled with an upper bound on the bias term, a result of this type gives a band of σ values for which $\hat{m}_n(\cdot,\sigma)$ converges uniformly over J to $m(\cdot)$.

Invoke Theorem 2.1 with ε fixed and $\gamma_n = \beta_n^d$, for a fixed Euclidean class with $PF^s < \infty$. To get the stated uniform convergence it suffices to have $n\beta_n^d/n^{1/s}$ increasing faster than $\log n$, that is,

$$n^{1-1/s}\,\beta_n^d/\log n \to \infty$$

This is weaker than the corresponding conditions imposed by Mack and Silverman (1982) (for $d = 1$ they required $n^\delta\beta_n \to \infty$ for some $\delta < 1 - 1/s$) or by Stute (1986) (the summability condition of his Theorem 3 implies that $n^\delta\beta_n^d/\log n \to \infty$ for some $\delta < 1 - 1/s$). Under the condition on $\{\beta_n\}$, we may express the assertions of the theorem for $\mathcal{F}$ and $\mathcal{G}$ as:

$$P_n f_{t,\sigma} = Pf_{t,\sigma} + o(P_n|f_{t,\sigma}| + P|f_{t,\sigma}| + \beta_n^d) \qquad \text{almost surely}$$

$$P_n g_{t,\sigma} = Pg_{t,\sigma} + o(P_n|g_{t,\sigma}| + P|g_{t,\sigma}| + \beta_n^d) \qquad \text{almost surely}$$

Here the $o(\cdot)$ terms should be interpreted to mean that the bound on the remainder terms holds uniformly in t and σ.

Restricting t to the set J and requiring $\sigma \geq \beta_n$, we get by division

$$\hat{m}_n(t,\sigma) = \frac{Pg_{t,\sigma} + o(\sigma^d)}{Pf_{t,\sigma} + o(\sigma^d)} \qquad \text{almost surely}$$

Again the $o(\cdot)$ terms should be interpreted to hold uniformly. The assumed behavior of $\sigma^{-d}Pf_{t,\sigma}$ reduces the last ratio to $m(t,\sigma) + o(1)$, which gives (4.4).

□

(4.5) Example

Nolan and Pollard (1986) sketched a new method for proving an optimality theorem for cross-validated kernel density estimation, a result first established under different conditions by Hall (1983) and Stone (1984). The method depended in part upon the application of two empirical process results that were not explicitly stated. The theorems from this paper fill the gap. As both applications

are similar, only one of them will be explained here. A complete discussion has appeared in Nolan (1986).

The problem concerns the choice of the smoothing parameter σ in the density estimation method of Example 4.1. For simplicity assume $p(\cdot)$ is a bounded density on the real line. Write $p(x,\sigma)$ for the density smoothed by a non-negative Euclidean kernel K.

At the end of Example 11 of Nolan and Pollard (1986), it was necessary to prove that $(P_n - P)(p(\cdot,\sigma) - p(\cdot))$ is eventually small, uniformly in σ, compared to

$$J_n(\sigma) = (n\sigma)^{-1} + \int [p(x,\sigma) - p(x)]^2 \, dx$$

That is, a uniform limit theorem was required for the class of functions g_σ defined, for $x \in \mathbb{R}$ and $\sigma > 0$, by

$$g_\sigma(x) = \int K\left(\frac{y - x}{\sigma}\right)[p(y) - p(x)] \, dy = \sigma[p(x,\sigma) - p(x)]$$

The desired result was

$$\sup_{\sigma>0} \frac{|P_n g_\sigma - P g_\sigma|}{\sigma \, J_n(\sigma)} \to 0 \qquad \text{almost surely} \tag{4.6}$$

Two applications of (2.5) provide the necessary justification.

Write $I(\sigma)$ for $\int [p(x,\sigma) - p(x)]^2 \, dx$. A simple application of Fatou's lemma shows that $I(\sigma)$ behaves like σ^4 as $\sigma \to 0$ and converges to a positive constant as $\sigma \to \infty$. It follows that the minimum of $J_n(\cdot)$ decreases no faster than $n^{-4/5}$. The connection between the bound (2.5) and (4.6) is the inequality

$$P g_\sigma^2 = \sigma^2 \int p(x)[p(x,\sigma) - p(x)]^2 \, dx \le C_1 \, \sigma^2 \, I(\sigma) \tag{4.7}$$

for some constant C_1.

Because $K(\cdot)$ is bounded, the smoothed density $p(x,\sigma)$ is uniformly of order $O(\sigma^{-1})$ as $\sigma \to \infty$. For a large enough constant C the contribution to (4.6) from those σ greater than C is bounded by

$$\frac{\varepsilon + |P_n p - Pp|}{\inf_{\sigma \geq C} J_n(\sigma)}$$

It is easy to dispose of the contributions from large σ .

On a bounded region $0 < \sigma \leq C$ the functions $g_\sigma(\cdot)$ are uniformly bounded. Rescale to make the bound 1, then invoke (2.5) to get, for $\lambda_n = n^{-1}\log n$,

$$\sup_{\sigma \leq C} \frac{|P_n g_\sigma - Pg_\sigma|}{(Pg_\sigma^2)^{1/2} + O(\lambda_n^{1/2})} = O(\lambda_n^{1/2}) \qquad \text{almost surely} \tag{4.8}$$

From (4.7),

$$\frac{(Pg_\sigma^2)^{1/2} + O(\lambda_n^{1/2})}{\sigma\, J_n(\sigma)}$$

$$\leq \frac{C_1^{1/2}\, \sigma\, I(\sigma)^{1/2} + O(\lambda_n^{1/2})}{\sigma((n\sigma)^{-1} + I(\sigma))}$$

$$\leq \frac{C_1^{1/2}}{((n\sigma)^{-1} + I(\sigma))^{1/2}} + \frac{O(n\lambda_n^{1/2})}{1 + n\sigma I(\sigma)}$$

$$= O(n^{2/5}) + O\left\{\frac{\log n}{1 + n\sigma^5}\right\} \lambda_n^{-1/2}$$

Let $\{\beta_n\}$ be a sequence of constants with $\lambda_n = o(\beta_n^5)$. On the region where $\sigma \geq \beta_n$, the last bound is uniformly of order $o(\lambda_n^{-1/2})$. Thus

$$\sup_{\beta_n \leq \sigma \leq C} \frac{|P_n g_\sigma - Pg_\sigma|}{\sigma\, J_n(\sigma)} = o(1) \qquad \text{almost surely}$$

The bound (4.8) is too wasteful for $\sigma < \beta_n$.

Because both $p(\cdot,\sigma)$ and $p(\cdot)$ are bounded, the functions $g_\sigma(\cdot)$ have a uniform $O(\beta_n)$ bound in the regions $0 < \sigma < \beta_n$. Rescale each g_σ by an appropriately large multiple of β_n ; then invoke (2.5) again to get

$$\sup_{\beta_n \leq C} \frac{|P_n g_\sigma - Pg_\sigma|}{(Pg_\sigma^2)^{1/2} + O(\beta_n \lambda_n^{1/2})} = O(\lambda_n^{1/2}) \qquad \text{almost surely}$$

Provided $\beta_n = o(1/\log n)$, which can be arranged without violating the constraint $\lambda_n = o(\beta_n^5)$, the extra factor of β_n is enough to handle the problem encountered with (4.8). That takes care of the contribution from $0 < \sigma < \beta_n$ to (4.6). □

§5. Proofs

The proofs depend on a symmetrization argument.

(5.1) Lemma

If $\{A(t) : t \in T\}$ and $\{B(t) : t \in T\}$ are independent permissible families of sets for which $\inf_t \mathbb{P}B(t) = \beta > 0$, then

$$\mathbb{P} \bigcup_t A(t) \le \beta^{-1} \mathbb{P}\, A(t)B(t) \qquad \square$$

For countable T this result is classical (Loeve 1977, Section 18.1.A). The treatment for uncountable T requires greater care. Independence must then be interpreted in terms of the coordinate projections of a product space. For the proofs that follow, this means that the auxiliary randomizations have to be constructed in a particular way. As the details of this construction are peripheral to the main arguments, further discussion is deferred to Section 6, where the general version of Lemma 5.1 is proved under more precisely stated conditions.

Proof of Theorem 2.1

As the asserted inequality is trivial for those ε_n greater than one, we may assume $0 < \varepsilon_n < 1$ for all n. Also we may assume that all members of every $\mathcal{F}_n$ are non-negative, since the general case would follow by separate consideration of positive and negative parts. And we may assume that g is continuous and strictly increasing, so that there is no ambiguity in the definition of $g^{-1}(n)$.

By the Strong Law of Large Numbers for $\{P_nF\}$, it is good enough to prove that

$$(5.2) \qquad \mathbb{P}\left\{\sup_{\mathcal{F}_n} \frac{|P_nf - Pf|}{P_nf + Pf + \gamma_n + \gamma_n P_nF + \gamma_n PF} > 8\varepsilon_n \text{ infinitely often}\right\} = 0$$

First we show that P_n can be replaced by a truncated process. For each f in $\mathcal{F}_n$ define $f_i(x) = f(x)\{g(F(x)) \le i\}$. Define T_n by

$$T_n f = n^{-1} \sum_{i=1}^{n} f_i(\xi_i)$$

Justify the substitution of T_n for P_n by proving

(5.3) $\sup_{\mathcal{F}_n} |T_n f - P_n f| = O(n^{-1})$ almost surely

(5.4) $|\mathbb{P}\, T_n f - Pf| \le \varepsilon_n Pf + \alpha_n$ for all f in $\mathcal{F}_n$

where $\{\alpha_n\}$ is a sequence of numbers of order $O(g^{-1}(n)/n\varepsilon_n)$.

Assertion (5.3) follows directly from the Borel-Cantelli lemma, because

$$\sum_{i=1}^{\infty} \mathbb{P}\{g(F(\xi_i)) > i\}$$

$$= \sum_{i=1}^{\infty} P\{g(F) > i\}$$

$$\le Pg(F)$$

$$< \infty$$

With probability one, only finitely many of the $\{\xi_i\}$ contribute to $\sup |T_n f - P_n f|$. [Note: this is the only place where the moment assumption (i) is used. For uniformly bounded classes, the truncation argument is unnecessary.]

For assertion (5.4) use $\alpha_n = PF\{g(F) > n\varepsilon_n\}$:

$$|\mathbb{P}\, T_n f - Pf|$$

$$\le n^{-1} \sum_{i=1}^{n} \mathbb{P}\, |f_i(\xi_i) - f(\xi_i)|$$

$$\le n^{-1} \sum_{i=1}^{n} P\, f\{g(F) > i\}$$

$$\le n^{-1}\, Pf\, (n \wedge g(F))$$

$$\le Pf\, \varepsilon_n + PF\{g(F) > n\varepsilon_n\}$$

On the set $\{g(F) > n\varepsilon_n\}$ assumption (ii) gives

$$g(F)/F \ge g(g^{-1}(n\varepsilon_n))/g^{-1}(n\varepsilon_n)$$

Because $\varepsilon_n < 1$ this implies

$$F \le g^{-1}(n)g(F)/n\varepsilon_n \qquad \text{on} \qquad \{g(F) > n\varepsilon_n\}$$

and so

$$\alpha_n \le (g^{-1}(n)/n\varepsilon_n)\ Pg(F)$$

The main condition of the theorem implies that $n\varepsilon_n^2\gamma_n/g^{-1}(n) \to \infty$. Thus both n^{-1} and $g^{-1}(n)/n\varepsilon_n$ are of order $o(\gamma_n\varepsilon_n)$. Assertion (4.1) will be a consequence of

$$(5.5) \qquad \mathbb{P}\left\{\sup_{\mathcal{F}_n} \frac{|T_nf - \mathbb{P}T_nf|}{P_nf + Pf + \gamma_n + \gamma_nP_nF + \gamma_nPF} > 6\varepsilon_n \quad \text{infinitely often}\right\} = 0$$

If we put

$$A_n(f) = \{|T_nf - \mathbb{P}T_nf| > 6\varepsilon_n(P_nf + Pf + \gamma_n + \gamma_nP_nF + \gamma_nPF)\}$$

then for (5.5) it will suffice to show

$$\sum_n \mathbb{P}(\cup_{\mathcal{F}_n} A_n(f)) < \infty$$

A symmetrization argument will give an appropriate bound for the n^{th} summand.

Let $\xi' = (\xi_1', \xi_2', \ldots)$ be a second sample from P, taken independently of $\xi = (\xi_1, \xi_2, \ldots)$. Construct T_n' and P_n' from the new sample. Define

$$B_n(f) = \{|T_n'f - \mathbb{P}T_n'f| \le \varepsilon_n(Pf + \gamma_n)\} \cap \{P_n'F < 2PF\}$$

By the Law of Large Numbers and Tchebychev's inequality,

$$\begin{aligned}
&\mathbb{P}B_n(f)^c \\
&\le \operatorname{var}(T_n'f) \,/\, \varepsilon_n^2(Pf + \gamma_n)^2 + o(1) \\
&\le (n\varepsilon_n)^{-2} \sum_{i=1}^{n} Pf_i^2 \,/\, (Pf + \gamma_n)^2 + o(1) \\
&\le (n\varepsilon_n)^{-2}\ nPf\ g^{-1}(n) \,/\, (Pf + \gamma_n)^2 + o(1) \\
&\le g^{-1}(n) \,/\, n\varepsilon_n^2\gamma_n + o(1) \\
&\longrightarrow \quad 0 \qquad \text{as} \qquad n \to \infty
\end{aligned}$$

Thus there exists an n_0 for which $\mathbb{P}\ B_n(f) \ge \frac{1}{2}$ for all f, if $n \ge n_0$. Lemma 5.1 gives

(5.6) $$\mathbb{P}(\cup_{\mathfrak{F}_n} A_n(f)) \le 2\,\mathbb{P}(\cup_{\mathfrak{F}_n} A_n(f)B_n(f)) \qquad \text{for } n \ge n_0 .$$

The following inequalities are implied by the three conditions that define the intersection $A_n(f)B_n(f)$:

$$|T_nf - T_n'f| > \varepsilon_n(6P_nf + 5Pf + 5\gamma_n + 6\gamma_nP_nF + 6\gamma_nPF)$$

$$6\gamma_n PF > 3\gamma_nP_n'F$$

$$\gamma_n + 2Pf > T_n'f$$

and consequently

$$|T_nf - T_n'f| > \varepsilon_n(6P_nf + 3Pf + T_n'f + 4\gamma_n + 6\gamma_nP_nF + 3\gamma_nP_n'F)$$

$$> \varepsilon_n(W_n(f) + W_n'(f))$$

where

$$W_n(f) = T_nf + 2\gamma_n + 2\gamma_nP_nF$$

$$W_n'(f) = T_n'f + 2\gamma_n + 2\gamma_nP_n'F$$

Introduce new independent sign variables $\sigma_1,\sigma_2,\ldots$ that are independent of ξ and ξ' and such that every σ_i takes the values $+1$ and -1 each with probability ½. Then, for each n, the distribution of the stochastic process

$$\{T_nf - T_n'f,\ T_nf + T_n'f,\ P_nF + P_n'F : f \in \mathfrak{F}_n\}$$

is the same as the distribution of the process

$$\left\{n^{-1}\sum_{i=1}^{n}\sigma_i(f_i(\xi_i) - f_i(\xi_i')),\ T_nf + T_n'f,\ P_nF + P_n'F : f \in \mathfrak{F}_n\right\}$$

It follows that eventually

(5.7) $$\mathbb{P}\cup_{\mathfrak{F}_n} A_n(f)$$

$$\le 2\mathbb{P}\left\{\exists f\in\mathfrak{F}_n : |n^{-1}\sum_{i=1}^{n}\sigma_i(f_i(\xi_i) - f_i(\xi_i'))| > \varepsilon_n(W_n(f) + W_n'(f))\right\}$$

$$\le 4\mathbb{P}\{\exists f\in\mathfrak{F}_n : |T_n^{\circ}f| > \varepsilon_nW_n(f)\}$$

where

$$T_n^{\circ}f = n^{-1}\sum_{i=1}^{n}\sigma_if_i(\xi_i)$$

Bound the last probability by working conditionally on ξ. That leaves only the randomness due to the $\{\sigma_i\}$; for the moment P_n and T_n are fixed measures. Invoke the definition of covering numbers to find a subclass $\mathcal{F}_n^*$, containing at most $M_n = N_1(\varepsilon_n\gamma_n, P_n, \mathcal{F}_n)$ members, such that for every f in $\mathcal{F}_n$ there is an f^* in $\mathcal{F}_n^*$ for which

$$P_n|f - f^*| \le \varepsilon_n\gamma_n P_nF$$

Because

$$|T_n^\circ f - T_n^\circ f^*| \le P_n|f - f^*|$$

and

$$T_nf \qquad \ge \quad T_nf^* - P_n|f - f^*|$$

the validity of the inequality

$$|T_n^\circ f| > \varepsilon_n W_n(f) = \varepsilon_n(T_nf + 2\gamma_n + 2\gamma_n P_nF)$$

would necessarily entail (remember: $\varepsilon_n < 1$)

$$|T_n^\circ f^*| > \varepsilon_n(T_nf^* + 2\gamma_n)$$

Thus

$$\begin{aligned}
&\mathbb{P}\{\exists f \in \mathcal{F}_n : |T_n^\circ f| > \varepsilon_n W_n(f) \mid \xi\} \\
&\le \mathbb{P}\{\exists f \in \mathcal{F}_n^* : |T_n^\circ f| > \varepsilon_n(T_nf + \gamma_n) \mid \xi\} \\
&\le 1 \wedge M_n \max_{\mathcal{F}_n^*} \mathbb{P}\{|T_n^\circ f| > \varepsilon_n(T_nf + \gamma_n) \mid \xi\}
\end{aligned}$$

The 1 comes from the trivial upper bound for all conditional probabilities. For each fixed f, Hoeffding's inequality (Pollard 1984, page 192) implies

$$\begin{aligned}
&\mathbb{P}\{|T_n^\circ f| > \varepsilon_n(T_nf + \gamma_n) \mid \xi\} \\
&\le 2\exp(-\tfrac{1}{2}\varepsilon_n^2(T_nf + \gamma_n)^2 \,/\, n^{-2}\sum_{i=1}^{n} f_i(\xi_i)^2) \\
&\le 2\exp(-\tfrac{1}{2}n\varepsilon_n^2(T_nf\ \gamma_n) \,/\, g^{-1}(n)T_nf)
\end{aligned}$$

Average out over the distribution of ξ to get

$$\begin{aligned}
&\mathbb{P}\{\exists f \in \mathcal{F}_n : |T_n^\circ f| > \varepsilon_n(T_nf + 2\gamma_n + 2\gamma_n P_nF)\} \\
&\le \mathbb{P}[1 \wedge 2M_n\exp(-n\varepsilon_n^2\gamma_n/g^{-1}(n))] \ .
\end{aligned}$$

An appeal to inequality (5.7) and the Borel-Cantelli lemma complete the proof. □

Proof of Corollary (2.2)

With probability one it is eventually true that

$$\sup_{\mathcal{F}_n(\gamma_n)} \frac{|P_n f - Pf|}{P_n f + Pf} \le 2C\varepsilon_n \le \tfrac{1}{2}$$

Since the condition on $2C\varepsilon_n$ then rules out the possibility $Pf = 0$, the ratio $R_n(f) = P_n f/Pf$ is eventually well defined. The inequality

$$|R_n(f) - 1| \le 2C\varepsilon_n(R_n(f) + 1)$$

implies

$$-4C\varepsilon_n \le R_n(f) - 1 \le 8C\varepsilon_n .$$ □

Proof of Corollary (2.3)

The class $\bar{\mathcal{F}}_n = \{f/\alpha_n : f \in \mathcal{F}_n\}$ has envelope 1 and the same covering numbers as $\mathcal{F}_n$. Argue as for the proof of Theorem 2.1 with $\bar{\mathcal{F}}_n$ instead of $\mathcal{F}_n$ and γ_n/α_n instead of γ_n. The truncation part of the argument is unnecessary this time; the role of $g^{-1}(n)$ is taken over by the constant 1. □

Proof of Theorem 2.4

There is no loss in generality to assume that $0 < \varepsilon_n < 1$ for all n. As for Theorem 2.1, there will be a reduction to a discrete problem by means of a symmetrization. As before P_n' will be the empirical measure constructed from the independent sample $\{\xi_i'\}$, and the $\{\sigma_i\}$ will be the independent sign variables. To simplify the notation write

$$\rho(f) = (Pf^2)^{1/2}$$

$$\rho_n(f) = (P_n f^2)^{1/2}$$

$$\rho_n'(f) = (P_n' f^2)^{1/2}$$

$$\rho_n^{\dagger}(f) = (P_n f^2 + P_n' f^2)^{1/2}$$

Notice that $\max(\rho_n(f),\rho_n'(f)) \le \rho_n^{\dagger}(f) \le \rho_n(f) + \rho_n'(f)$.

For each f in $\mathcal{F}_n$ define events

$$A_n(f) = \{|P_n f - Pf| > 9\varepsilon_n(\rho_n(f) + \rho(f) + \gamma_n\rho_n(F) + \gamma_n\rho(F))\}$$

$$B_n(f) = \{|P_n' f - Pf| \le \varepsilon_n\rho(f)\} \cap \{\rho_n'(f) \le 2\rho(f)\} \cap \{\rho_n'(F) \le 2\rho(F)\}$$

It is enough to prove that

$$\sum_n \mathbb{P}(\cup_{\mathcal{F}_n} A_n(f)) < \infty$$

Bound the n^{th} summand by an appeal to Lemma 5.1. First show that $\inf \mathbb{P}B_n(f) \ge \frac{3}{4}$ for all n large enough.

$$\mathbb{P}B_n(f)^c \le \operatorname{var}(P_n'f)/\varepsilon_n^2\rho(f)^2 + \mathbb{P}\rho_n'(f)^2/4\rho(f)^2 + \mathbb{P}\{\rho_n'(F) > 2\rho(F)\}$$

The first term on the righthand side is less than $(n\varepsilon_n^2)^{-1}$, which converges to zero by virtue of the summability condition of the theorem; the second term equals $\frac{1}{4}$; the third term converges to zero by virtue of a law of large numbers.

On the intersection of the sets $A_n(f)$ and $B_n(f)$ we have

$$|P_n f - P_n' f| > \varepsilon_n(9\,\rho_n(f) + 4\rho_n'(f) + 9\gamma_n\rho_n(F) + 4\tfrac{1}{2}\gamma_n\rho_n'(F))$$

$$> 4\varepsilon_n(\rho_n^+(f) + \gamma_n\rho_n^+(F))$$

Thus Lemma 5.1 gives, for all n large enough,

$$\mathbb{P}(\cup_{\mathcal{F}_n} A_n(f)) \le 2\mathbb{P}\{\exists f \in \mathcal{F}_n : |P_n f - P_n' f| > 4\varepsilon_n(\rho_n^+(f) + \gamma_n\rho_n^+(F))\}$$

Notice the symmetry in the probability expression on the righthand side; it would be unaffected by an interchange of any (ξ_i,ξ_i') pair. It would also be unaffected by the random interchange of pairs induced by the sign variables $\{\sigma_i\}$, as in the proof of Theorem 2.1. If P_n° denotes the symmetrized empirical measure for which

$$P_n^\circ f = n^{-1} \sum_{i=1}^n \sigma_i f(\xi_i),$$

then the last probability is less than

$$\mathbb{P}\{\exists f \in \mathcal{F}_n : |P_n^\circ f| > 2\varepsilon_n(\rho_n(f) + \gamma_n\rho_n(F))\}$$

Bound this probability by working conditionally on ξ.

Given ξ, find a subclass $\mathcal{F}_n^*$ containing at most $M_n = N_2(\gamma_n, P_n, \mathcal{F}_n)$ members, such that for every f in $\mathcal{F}_n$ there is an f^* in $\mathcal{F}_n^*$ for which

$$\rho_n(f - f^*) \leq \varepsilon_n \gamma_n \rho_n(F)$$

Because

$$|P_n^\circ f - P_n^\circ f^*| \leq P_n|f - f^*| \leq \rho_n(f - f^*)$$

and (remember: $\varepsilon_n < 1$)

$$\rho_n(f^*) \geq \rho_n(f) - \gamma_n \rho_n(F)$$

the validity of the inequality

$$|P_n^\circ f| > 2\varepsilon_n(\rho_n(f) + \gamma_n \rho_n(F))$$

would necessarily entail

$$|P_n^\circ f^*| > 2\varepsilon_n \rho_n(f^*)$$

Thus

$$\begin{aligned}
&\mathbb{P}\{\exists f \in \mathcal{F}_n : |P_n^\circ f| > 2\varepsilon_n(\rho_n(f) + \gamma_n \rho_n(F)) \mid \xi\} \\
&\leq \mathbb{P}\{\exists f \in \mathcal{F}_n^* : |P_n^\circ f| > 2\varepsilon_n \rho_n(f) \mid \xi\} \\
&\leq 1 \wedge M_n \max_{\mathcal{F}_n^*} \mathbb{P}\{|P_n^\circ f| > 2\varepsilon_n \rho_n(f) \mid \xi\} \\
&\leq 1 \wedge M_n \, 2 \exp(-2n\varepsilon_n^2 \rho_n(f)^2 / \rho_n(f)^2) \,,
\end{aligned}$$

the last bound coming from Hoeffding's inequality. Notice how the $\rho_n(f)$ is exactly the right weighting for this inequality.

Average out over the distribution of ξ, then sum over n to complete the proof. □

§6. Permissibility

Appendix C of Pollard (1984) defines an indexed class of functions $\{f(\cdot,t) : t \in T\}$ to be permissible if each $f(\cdot,\cdot)$ is measurable as a function on a product space, and T is a Souslin measurable space. This definition is adequate for the usual sort of function class indexed by a finite dimensional parametrization. It takes care of most of the measure-theoretic technicalities that arise in the symmetrization arguments of this paper. But, as a trivial example shows, permissibility alone is not enough for a precise statement of Lemma 5.1.

Let Ω be the unit interval equipped with Lebesgue measure $\mathbb{P}$ on its Borel σ-field. For each t in $[0,1]$ define $A(t)$ as the singleton $\{t\}$, and $B(t)$ as the complement of $A(t)$. The σ-field $\mathcal{A}$ generated by $\{A(t) : 0 \le t \le 1\}$ contains only countable sets and their complements. Each $B(t)$ has probability 1 and is independent of $\mathcal{A}$. But nevertheless

$$\mathbb{P} \bigcup_t A(t)B(t) = 0 < 1 = \mathbb{P} \bigcup_t A(t)$$

Clearly a definition of independence based only on finite dimensional distributions is inadequate for the families of sets in Lemma 5.1. Here is a better version.

(6.1) Lemma

Let $\Omega \otimes \Omega'$ be a product space equipped with a product σ-field $\mathcal{L} \otimes \mathcal{L}'$ and a product probability measure $\mathbb{P} \otimes \mathbb{P}'$. Let $\{A(t) : t \in T\}$ be a permissible family of subsets of Ω and $\{B(t) : t \in T\}$ be a permissible family of subsets of $\Omega \otimes \Omega'$. If there exists a $\beta > 0$ for which

$$\inf_t \mathbb{P}' B(t) \ge \beta$$

then

$$\mathbb{P} \bigcup_t A(t) \le \beta^{-1} \mathbb{P} \otimes \mathbb{P}' [\bigcup_t (A(t) \otimes \Omega') \cap (\Omega \otimes B(t))]$$

Proof. Write A for the set $\{(t,\omega) : \omega \in A(t)\}$. By definition it is a measurable subset of $T \otimes \Omega$. The cross-section theorem for Souslin spaces (Dellacherie and Meyer (1978), III.45) gives a measurable map τ from Ω into $T \cup \{\infty\}$ such that

$$(\tau(\omega),\omega) \in A \qquad \text{whenever} \qquad \tau(\omega) \neq \infty$$

and

$$\tau(\omega) \neq \infty \quad \text{for } \mathbb{P} \text{ almost all } \omega \text{ in } \bigcup_t A(t)$$

Thus

$$\begin{aligned}
\mathbb{P} \bigcup_t A(t) &\leq \mathbb{P}\{\omega : \tau(\omega) \neq \infty\} \\
&\leq \beta^{-1}\, \mathbb{P}\{\omega : \tau(\omega) \neq \infty\}\, \mathbb{P}'\{\omega' : \omega' \in B(\tau(\omega))\} \\
&\leq \beta^{-1}\, \mathbb{P} \otimes \mathbb{P}'\{(\omega,\omega') : \omega \in A(\tau(\omega)),\ \omega' \in B(\tau(\omega))\} \\
&\leq \beta^{-1}\, \mathbb{P} \otimes \mathbb{P}'\left[\bigcup_t (A(t) \otimes \Omega') \cap (\Omega \otimes B(t))\right]
\end{aligned}$$

□

REFERENCES

Alexander, K.S. (1984). Rates of growth for weighted empirical processes, in Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, L. Le Cam and R. Olshen, eds. Wadsworth, Belmont, California.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and Regression Trees. Wadsworth, Belmont, California.

Dellacherie, C. and Meyer, P-A. (1978). Probabilities and Potential, Part A. North-Holland, Amsterdam.

Dudley, R.M. (1978). Central limit theorems for empirical measures. Annals of Probability 6, 899-929. (Correction, ibid. 7 (1979), 909-911.)

Dudley, R.M. (1986). Universal Donsker classes and metric entropy. Annals of Probability (to appear).

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. Annals of Statistics 11, 1156-1174.

Le Cam, L. (1983). A remark on empirical measures, in Festschrift for E.L. Lehmann, eds. P. Bickel, K. Doksum, and J. Hodges. Wadsworth, Belmont, California.

Loeve, M. (1978). Probability Theory, II (Fourth edition). Springer-Verlag, New York.

Mack, Y.P. and Silverman, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. Z. Wahrscheinlichkeitstheorie verw. Geb. 61, 405-415.

Nolan, D. (1986). U-processes. Ph.D. dissertation, Yale University.

Nolan, D. and Pollard, D. (1986). U-processes: rates of convergence. Annals of Statistics (to appear).

Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Annals of Statistics 12, 1285-1297.

Stute, W. (1986). On almost sure convergence of conditional empirical distribution functions. Annals of Probability 14, 891-901.